

**Artificial Neural Networks: Breast Cancer Classification Report**

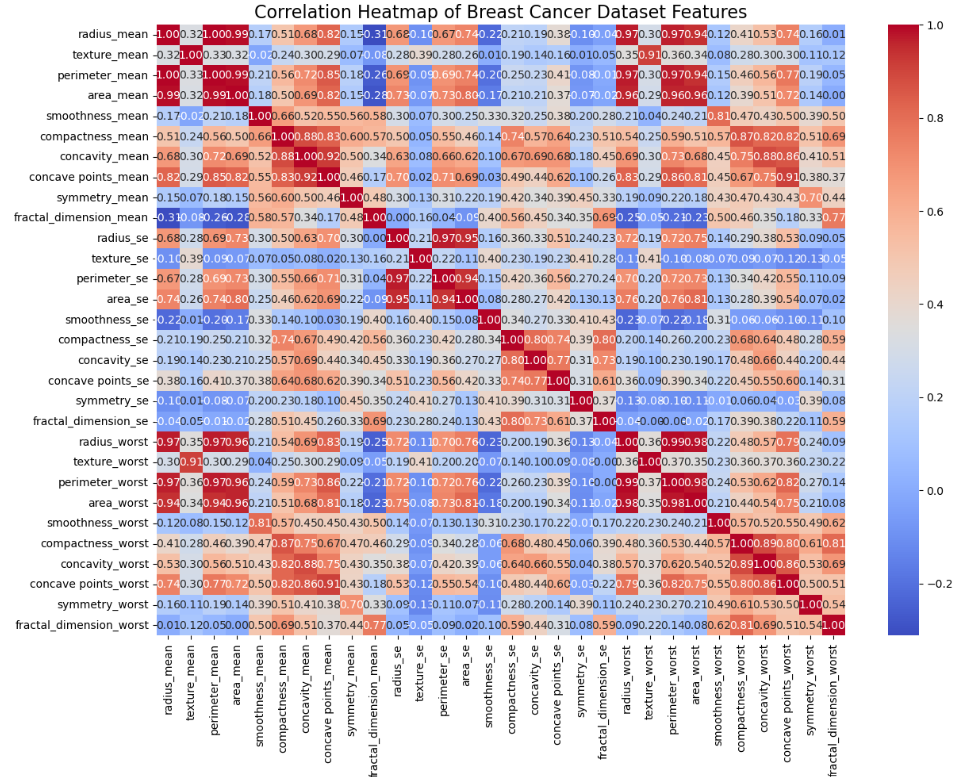
## Breast Cancer: Analysis and Classification

**Introduction:**

Breast cancer is a major health concern around the world, accounting for 25% of all cancer incidences among women. With over 2.1 million afflicted persons in 2015 alone, effectively detecting and categorizing breast cancers is critical for timely intervention and treatment. In this project, we will build, refine, and evaluate a Multilayer Perceptron (MLP) classifier for breast cancer diagnosis, using a well-curated dataset of X-ray images to extract relevant characteristics. Our key goals include feature selection, data normalization, model creation, and comprehensive evaluation, all of which aim to improve the MLP's performance in distinguishing between malignant and benign tumors.

**Data collection and preprocessing:**

The dataset for this analysis includes a variety of features collected from breast tumor X-ray pictures, as well as the corresponding diagnosis (malignant or benign). To assist with feature selection, we separated the dataset into features and target variables after importing it.



The above heatmap will show the correlation coefficients for each pair of attributes. Positive correlations are depicted in warmer hues (closer to red), whilst negative correlations are portrayed in cooler colors (near to blue). Values near 1 or -1 suggest stronger correlations, whereas values near 0 indicate weaker correlations.

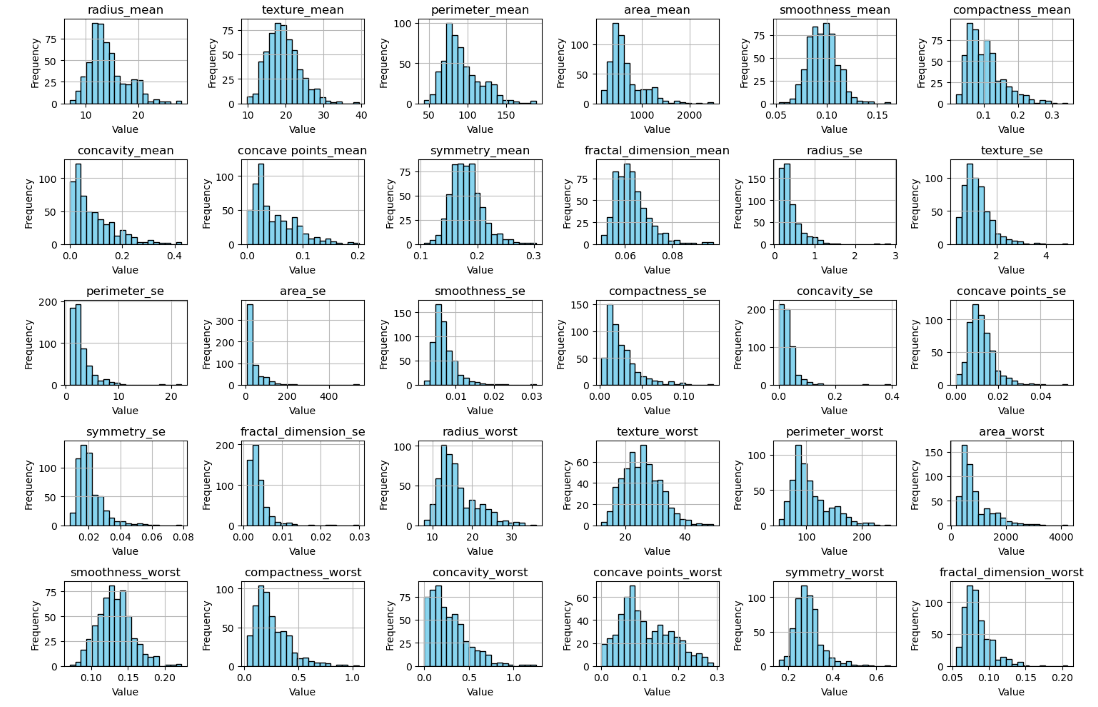
A correlation value around 1 indicates a significant positive linear relationship between the two variables, meaning that as one rises, the other also tends to rise. Conversely, a correlation value of about -1 indicates a strong negative linear relationship, in which one variable tends to decline as the other grows.

Correlation coefficients around zero, on the other hand, indicate a weak or insufficient linear link between the variables. This indicates that there is no consistent relationship between changes in one variable and changes in the other. Lighter shades in the heatmap indicate weaker correlations; values closer to 0 are reflected by these tones. With a short glance at the heatmap's colors and values, analysts can spot patterns of relationship between various

Moving on to Normalization, Several factors influence the decision to choose **MinMaxScaler()** scaling as the normalization approach, including the properties of the dataset and the classification task requirements. This approach preserves the relationships between original data points, which is critical for features with different scales, guaranteeing that no single characteristic dominates others. Min-Max scaling is especially useful for Multilayer Perceptron (MLP) classifiers, which are sensitive to input feature scales. It stabilizes training, reduces gradient difficulties, and accelerates convergence. Furthermore, it effectively tackles outliers by compressing the data range(in our case [0,1]), maintains interpretability, and provides simplicity and universality in implementation across various datasets and methods. Overall, these characteristics help to increase model performance and interpretability in the breast cancer classification job.

**Features Selection :**

(The below histograms showcase each feature to visualize their distributions. This will help us understand each feature's range and spread of values.)



Feature selection is an important stage in developing effective machine-learning models. For this work, we used the chi-squared test within the SelectKBest approach to select the top six traits from all the above features strongly connected with breast cancer diagnosis. The selected features provide important insights into tumor properties required for effective classification.

**radius\_mean**: The mean of distances from the center to points on the perimeter.

**texture\_mean**: Standard deviation of gray-scale values.

**perimeter\_mean**: Sum of the lengths of segments forming the boundary.

**area\_mean**: Area of the tumor.

**smoothness\_mean**: Variation in the radius lengths.

**compactness\_mean**: Measure of compactness calculated as perimeter^2 / area - 1.0.

**SelectKBest** was chosen for feature selection in this breast cancer classification job based on the chi-squared test because it works well with both categorical target variables and a combination of numerical and categorical input characteristics. This strategy reduces dimensionality while retaining the most useful features by ranking features according to their relevance in predicting the target variable.

The chi-square test is essentially a statistical method that can be used for assessing if there is a significant association between category variables. To confirm if the fluctuations are the result of chance or a relationship between the variables, it compares the observed and anticipated frequencies of the categories. It is frequently used to assess data and make inferences about the correlations between variables in fields including biology, sociology, and market research.

By selecting a smaller selection of characteristics, the model improves overall interpretability, decreases overfitting, and increases computational efficiency. Because of this, it's a practical and effective way to identify the most relevant features in the dataset.

**Model Building:**

A neural network model called the Multilayer Perceptron Classifier, or MLPClassifier(), is employed to solve classification issues. It consists of an input layer, one or more hidden levels, and an output layer, among other layers of nodes. Every network node is connected to every other node in the adjacent layers, and each connection has a weight assigned to it. To minimize a predefined loss function, the model modifies these weights during training based on the input data, frequently using gradient descent and backpropagation. MLPClassifier is frequently used in machine learning applications like image recognition, natural language processing, and medical diagnosis because of its ability to identify complex patterns in data.

The model, MLPClassifier, is configured as follows:

Hidden Layer Configuration: It has a single hidden layer with 100 neurons. This choice seeks to create a compromise between model complexity and computing efficiency.

The Rectified Linear Unit (ReLU) activation function is used. ReLU is often used because of its ability to reduce the vanishing gradient problem and accelerate convergence.

Optimizer: The Adam optimizer is used to train the model. Adam is well-suited for huge datasets and offers flexible learning rates, making it efficient and effective at optimizing model parameters.

Regularization: L2 regularization with an alpha value of 0.0001 is used to prevent overfitting and improve the model's generalization ability.

A batch size of 32 is used during training. This batch size achieves a compromise between computing efficiency and model convergence.

The starting learning rate is set at 0.001. This value defines the step size during optimization, which influences convergence speed and stability.

Maximum Iterations: For training purposes, 200 iterations is the maximum. By restricting the number of iterations, this parameter controls the training duration and aids in the prevention of overfitting.

Random State: For reproducibility, a random state of 42 is used, guaranteeing that the results are constant across runs.

The explanation explains why certain hyperparameters and settings were chosen for the MLPClassifier model. Each option is carefully justified based on its relevance to the classification challenge and previous success in neural network modeling. Every decision, from the single hidden layer with 100 neurons to the use of the ReLU activation function, Adam optimizer, L2 regularization, and other parameters like batch size, learning rate, maximum iterations, and random state, seeks to strike a balance between model complexity, computational efficiency, regularization, and reproducibility. Together, these considerations contribute to a well-optimized MLPClassifier model that can successfully capture complicated patterns in data while avoiding overfitting and providing consistent and predictable performance across multiple runs.

**Results & Evaluation:**

The confusion matrix gave a breakdown of the model's predictions:

True Positives (TP): The model correctly identified 35 cases as malignant tumors.

True Negatives (TN): The model correctly identified 71 cases as benign tumors.

False Positives (FP): The model incorrectly identified 0 cases of benign tumors as malignant.

False Negatives (FN): The model incorrectly identified 8 cases of malignant tumors as benign

Accuracy: Accuracy assesses the overall correctness of the classifier's predictions. In this situation, the model had an accuracy of about 95.56%. This means that around 92.98% of the model's predictions were accurate.

Precision: Precision is the fraction of true positive forecasts to all positive predictions. In this case, precision represents the model's capacity to prevent false positives. A precision score of 0.95 implies that all of the model's positive predictions were accurate.

Recall: Recall, also known as sensitivity, calculates the fraction of accurate positive predictions among all positive instances. In this situation, the model had an approximate recall of 93%. This indicates that the model accurately recognized approximately 81.40% of the malignant tumors in the dataset.

F1 Score: The F1 score represents the harmonic mean of precision and recall. It provides a balanced evaluation of the model's performance, taking into account both precision and recall. In this case, the F1 score is at 94.%, showing a solid mix of precision and recall.

Given the severe consequences of misdiagnosis (misclassifying malignant tumors as benign), avoiding false negatives is one of our top priorities for choosing the right metrics for classifying malignant and benign tumors. Recall (sensitivity), which explains the model's accuracy in identifying malignant tumors, thus becomes the key measure. Although it comes in second to recall, precision is evaluated as well because it shows how well the model classifies malignant tumors and can be used to gauge the accuracy of positive predictions. Since the F1 score provides a thorough assessment of the model's performance and strikes a balance between precision and recall, it is considered significant. Recall, precision, and the F1 score together offer a comprehensive evaluation that emphasizes how crucial it is to correctly detect malignant breast cancer.

Overall, The evaluation of the MLP classifier demonstrated decent performance across a variety of parameters. These findings support the MLP classifier's ability to accurately classify malignant and benign tumors.